Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 11, No. 2, 2023

Random Forest-Based Restocking and Pricing Prediction for Vegetable Items

Kaile Wang[1, †], Keming Su[2, †], Hao Li[2, †]

[1] School of Information Engineering, Inner Mongolia University of Science & Technology, Baotou, China
[2] School of Civil Engineering, Inner Mongolia University of Science & Technology, Baotou, China
[†] These authors also contributed equally to this work

Abstract: The intolerance to storage of vegetable commodities in supermarkets makes automatic pricing and replenishment decisions for vegetable commodities particularly important. This paper takes the measured data of a superstore as an example to formulate a set of effective pricing and replenishment decisions for vegetable commodities, which is a comprehensive consideration to ensure the balance of supply and demand, and to reduce the losses of the superstore and the loss rate of commodities. First of all, the sales of each category and single product in different time periods were counted, and the Pearson correlation coefficient was calculated to obtain the distribution pattern of the sales volume of each category and single product of vegetables. Then, the relationship between the total sales volume of and the cost-plus pricing of each vegetable category is analyzed, and a random forest model is established to predict the total replenishment volume and pricing strategy in the coming week. Finally, the replenishment quantity and pricing strategy of individual items are given to maximize the revenue of the superstore under the premise of trying to meet the market demand for each category of vegetable goods. The model established in the paper, which basically solves the given problem, has strong practicality and high computational efficiency.

Keywords: Pricing Strategies, Pearson Correlation Coefficient, Regression, Random Forest.

1. Introduction

In recent years, with the development of our country, the market demand for fresh products is also expanding. However, the problem has also emerged, in China's fresh food industry chain, it will be due to the mass demand, supply capacity, price discomfort and other reasons to cause problems such as stagnant sales of fresh food, fresh food in short supply, and cause loss of revenue and waste of commodities and other problems [1-3]. The reason for this phenomenon is that the sales strategy is not suitable, the supply and demand is not enough to understand, and the pricing mechanism is unreasonable. At present, most of the domestic supermarket fresh sales, the whole process of fresh sales for the price of the same, the approach of the low price sold at the end of the period, resulting in a large number of fresh rot and stagnation, not only let the supermarket and fruit and vegetable dealers to suffer serious economic losses, but also make the fresh to cause unnecessary waste.

With the improvement of people's living standards and the rapid dissemination of information, consumers show diversity in their purchasing decisions, focusing not only on the value and cost-effectiveness of goods, but also on the quality and safety of goods. The intolerance to storage of vegetable products in supermarkets makes automatic pricing and replenishment decisions for vegetable products particularly important, which involve the influence of many factors such as market demand, supply, and cost. In order to ensure effective pricing and replenishment decisions for vegetables, as well as to improve the shopping experience of customers, we need to take into account to ensure the balance between supply and demand, so as to reduce the loss of superstores and the loss rate of goods [4-6].

In order to solve the above problems, this paper takes the measured data of a superstore as an example, firstly, to find out the distribution law of the sales volume of each vegetable category and single product and their interrelationship. Then, analyze the relationship between the total sales volume of each vegetable category and the cost-plus pricing, and give the total daily replenishment and pricing strategy of each vegetable category in the coming week, so that the superstore can maximize the revenue. Finally, due to the limited sales space of vegetable items in the superstore, a replenishment plan for new individual items is formulated to maximize the revenue of the superstore under the premise of trying to satisfy the market's demand for vegetable items in each category.

2. Sales Volume Distribution Pattern and Correlation Analysis

The measured data of a hypermarket gives the categories to which all vegetables in the hypermarket belong and the sales flow of each individual item. Analyzing and processing the measured data of a superstore, we get the sales of each category and single product in different quarters . According to the sales situation in different time periods, the distribution pattern of the sales volume of each category and single product of vegetables can be inferred.

The change in sales (in kilograms) of individual dishes from the third quarter of 2020 to the second quarter of 2023 is shown in Table.1.

Table 1. Statistics on total sales by category in each quarter

kind season	philodendron	cauliflower	Aquatic rhizomes	eggplant	capsicum	edible fungi
2020 Q3	6589.717	2535.512	7281.936	2984.788	7012.806	6149.678
2020 Q4	2314.155	2919.946	9250.989	1354.473	5174.008	11643.83
2021 Q1	3642.985	2059.854	9853.369	1931.269	8138.264	11004.17
2021 Q2	6124.785	2172.837	6411.767	2010.647	6187.02	4621.28
2021 Q3	6334.069	2155.516	8315.872	2513.558	4888.864	5101.016
2021 Q4	2260.045	2035.208	9391.983	913.862	3102.582	7343.637
2022 Q1	2293.363	2746.059	9149.283	1707.858	5463.779	5562.767
2022 Q2	4578.351	1890.901	6234.736	2576.898	1556.909	3352.926
2022 Q3	2160.394	3527.74	9149.877	1299.507	7760.977	6499.568
2022 Q4	474.35	2129.833	10489.672	623.969	7931.845	11748.3
2023 Q1	740.752	2026.572	9181.67	1438.505	9550.94	10947.53
2023 Q2	2558.88	1338.122	6724.201	2028.759	6222.956	6850.968

In order to more graphically represent the volume of sales in each category for each season, the seasonal sales volume is plotted as shown in Figure 1.

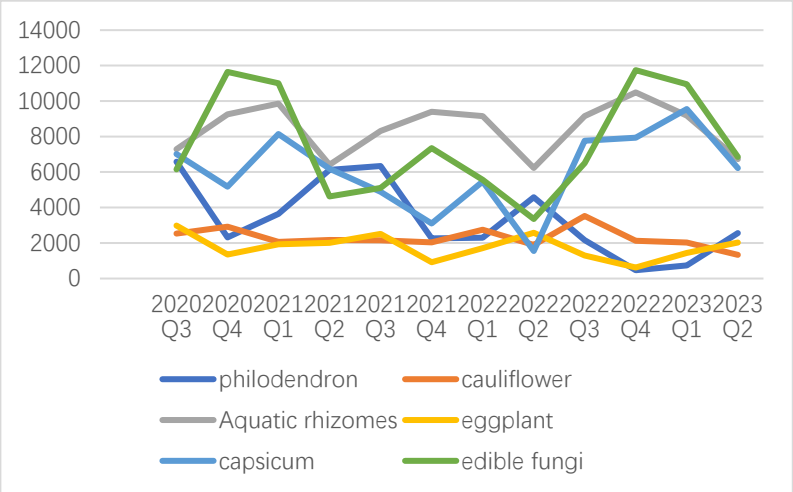

Figure 1. Quarterly sales of six major types of vegetables

Assuming that the sales volume of the two highest-selling individual products in each category of vegetables reflects the distribution pattern and interrelationship of the sales volume of individual products in that category, and because it is more complicated to write out the interrelationship and distribution pattern of the sales volume of all the vegetable categories in the text, we use eggplant as a representative.

The two highest selling items in the eggplant category are: purple eggplant (2) and green eggplant (1), and the line graph of sales of these two eggplants by quarter is shown in Figure 2.

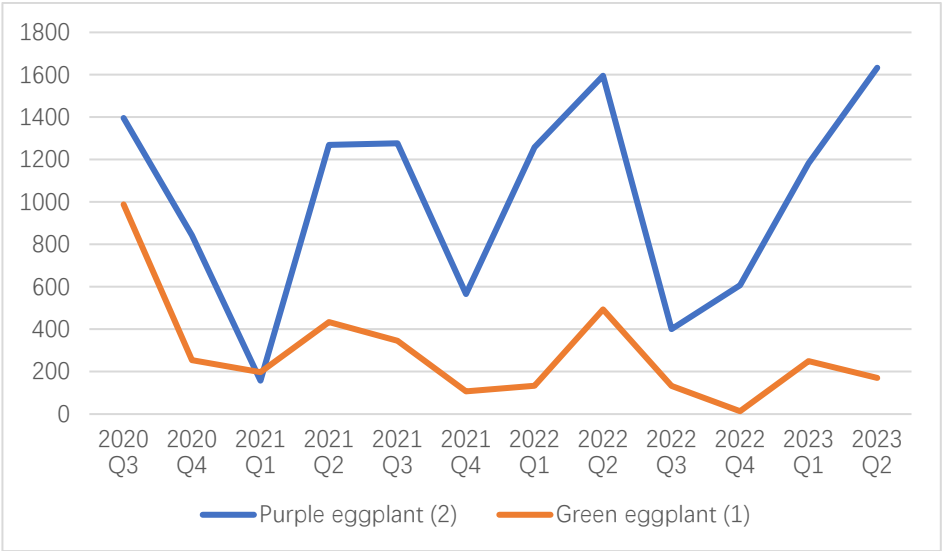

Figure 2. Line graph of sales of purple eggplant (2) and green eggplant (1) by quarter

There are two types of eggplant in each quarter after the sales, according to the Pearson correlation coefficient calculation process, the specific Pearson correlation coefficient calculation formula is as follows:

$$R = \frac{N\sum x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{N\Sigma x_i^2 - (\sum x_i)^2}\sqrt{N\Sigma y_i^2 - (\Sigma y_i)^2}} \quad (1)$$

The correlation coefficient was calculated using MATLAB: 0.5009.

The associated strength rating scale is shown in Table.2. Since R=0.5009 < 0.6 , it indicates that the positive correlation between the top two selling items in the eggplant category is moderately correlated.

Table 2. Table of relevant strength classes

Numerical range	degree of relevance
0.8-1.0	Highly relevant
0.6-0.8	strong correlation
0.4-0.6	Moderately relevant
0.2-0.4	weak correlation
0.0-0.2	Very weak correlation or no correlation

3. Individual Product Replenishment Program Development

The sales volume and selling price of each day of the vegetable category were obtained, and then the data of the first day of the latter 18 months were selected to establish a regression model, so as to derive the relationship between the total sales volume and the average unit price of each vegetable category.

It is known that the total daily replenishment is positively correlated with the total sales volume, so the pricing strategy is determined by the positive correlation between the ratio of wholesale unit price and sales unit price. Since the variation of daily sales volume and wholesale price is small, we selected the wholesale price and sales volume in the recent week and predicted the sales volume in the coming week by using the random forest model.

Data analysis was carried out to obtain the values of sales and average selling price of cauliflower category as shown in Table 3.

Table 3. Cauliflower sales and average unit price

sales volume	Average unit price
62	8
13.968	10
22.944	10
25.701	8.4
23.295	10
22.234	8
28.897	8.9
82.828	8.4
47.503	12.7
76.779	10.1
39.603	7.8
24.722	5
39.38	7.2
33.307	9.6
23.98	8.5
35.079	10.7
29.389	10.2
19.364	13.7

To make the relationship more obvious we made a line graph of the two, as shown in Figure 3.

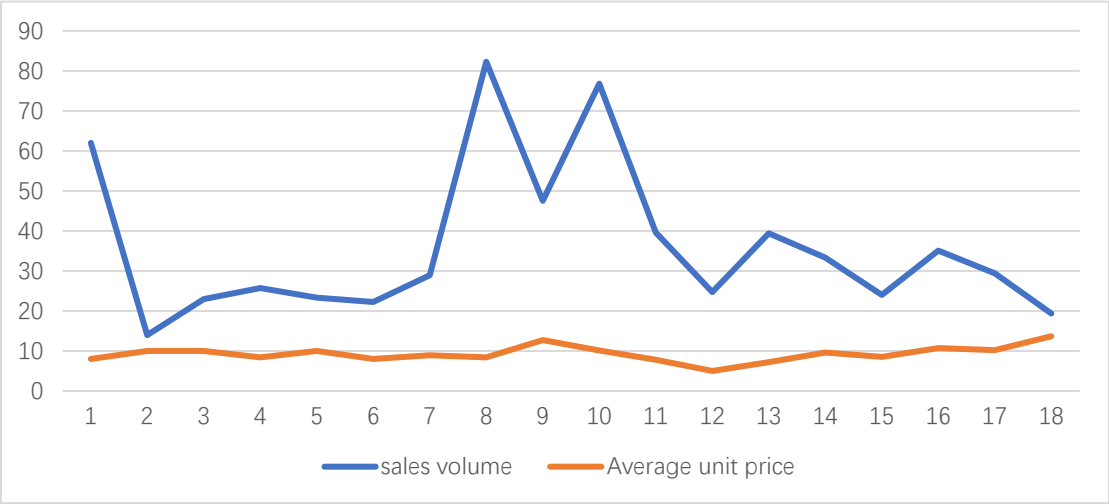

Figure 3. Relationship between sales volume and average unit price of cauliflower category

Through the line graph we can roughly see that there is a certain relationship between the two, so a regression model is established for correlation analysis:

$$\begin{cases} y_{i\_cabbage} = \beta_0 + \beta x_{i\_cabbage} \\ \xi \sim N(0, \sigma^2) \end{cases} \quad (2)$$

Eq.ξ obeys a normal distribution, withβ_0 is the constant term coefficient and β is the coefficient of the independent variable $y_{i\_cabbage}$ denotes the price of cauliflower on the first i day cauliflower pricing $x_{i\_cabbage}$ denotes the price of cauliflower on day i We imported the obtained data into MATLAB and calculated the Pearson's correlation coefficient of the two and finally came up with the result of R=0.5009.

The overall regression coefficient is not zero i.e. there is a linear relationship between the variables. Significant p is 0.0004, which presents significance at the level, so the model basically meets the requirements finally the relationship is

obtained as follows:

$$y_{i\_cabbage}=10.2865-0.0466x_{i\_cabbage} \quad (3)$$

In order to visualize the change between the two more the regression equation was fitted to the original data observed as shown in Figure 4.

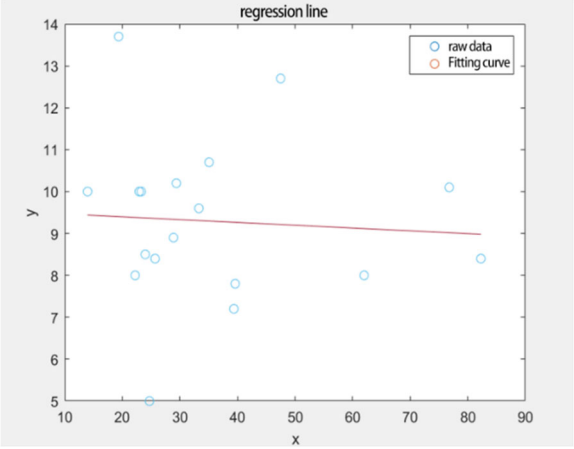

Figure 4. Regression equation fitted to the original data

In this question we mainly use the random forest model to make predictions about the total daily replenishment and pricing strategies for the coming week. The random forest model has the following characteristics:(1) It is able to deal with high-dimensional data and a large number of features without the need for feature selection. (2) Capable of estimating feature importance during the training process to help understand the key features of the data. (3) Able to handle missing data and maintain good performance. (5) Has good robustness and can handle noise and outliers.

Based on the total sales volume of each vegetable category from June 5-30, 2023, the total replenishment volume for June 5-30, 2023 will be Based on the Random Forest model to predict the sales volume of each category for the next week, the model of Cauliflower category is used as an example to predict the replenishment volume of the Cauliflower category for the coming week by applying the Random Forest model. According to the random forest model, the total sales of each vegetable category for the next seven days can be obtained, as shown in Table.4.

Table 4. Forecasted sales volume

	cauliflower	philodendron	capsicum	eggplant wholesale	edible mushroom	Aquatic rhizomes
2023/7/1	16.1077	134.5336	100.8684	89.007	81.1291	19.3111
2023/7/2	16.071	133.3179	101.1612	91.3433	82.0071	19.4101
2023/7/3	16.0344	132.1021	101.454	93.6797	82.8851	19.5091
2023/7/4	15.9978	130.8863	101.7469	96.0161	83.7631	19.6081
2023/7/5	15.9612	129.6706	102.0397	98.3524	84.6411	19.7071
2023/7/6	15.9245	128.4548	102.3325	100.6888	85.5191	19.8061
2023/7/7	15.8879	127.239	102.6254	103.0252	86.3971	19.9051

Based on the ratio of wholesale price to sales price from July 1-7, 2023, the pricing strategy for 1-7, 2023 is derived, and the pricing strategy for the next seven days is calculated by the random forest model, as shown in Table.5.

Table 5. Forecast Pricing Strategy

	cauliflower	philodendron	capsicum	eggplant wholesale	edible mushroom	Aquatic rhizomes
2023/7/1	0.6212	0.6089	0.5086	0.5247	0.4503	0.4751
2023/7/2	0.6286	0.594	0.4992	0.5032	0.4256	0.4334
2023/7/3	0.636	0.5791	0.4899	0.4816	0.4009	0.3917
2023/7/4	0.6433	0.5643	0.4805	0.4601	0.3762	0.3499
2023/7/5	0.6507	0.5494	0.4712	0.4385	0.3515	0.3082
2023/7/6	0.6581	0.5346	0.4619	0.4169	0.3268	0.2665
2023/7/7	0.6665	0.5197	0.4525	0.3954	0.3021	0.2247

4. Individual Product Replenishment Strategy Development

Screening of a supermarket's measured data to obtain 36 saleable varieties that meet the requirements, and then through the profit formula and the loss rate formula comprehensive analysis of these 36 varieties, and finally determine the 33 saleable varieties sold on July 1st. By obtaining the replenishment volume and pricing strategy of individual items on July 1, the 33 sellable varieties are finally selected to maximize the revenue of the superstore.

The predicted single item replenishment for July 1 is obtained, as shown in Table.6.

Table 6. Individual product replenishment on July 1

cauliflower	philodendron	capsicum	eggplant wholesale	edible mushroom	Aquatic rhizomes
16.1077	134.5336	100.8684	89.007	81.1291	19.3111

By filtering the data, we can come up with 36 sellable varieties that fit the question , then calculate the profit and

attrition rate for these 36 sellable varieties and select 33 sellable individual items with the following profit formula:

$$l = (p - j)x \tag{4}$$

where l represents the profit of a single product, and p represents the wholesale price of a single product, and j represents the purchase price of a single product, and x represents the sales volume of a single product. Profit formula is as follows:

$$\mathrm{B} = \begin{cases} 1, & S > 9.43\% \\ -\frac{S}{S_m} + 2, & S \leq 9.43\% \end{cases} \tag{5}$$

where the wastage rate is S, the average wastage rate of the sample is S_m , B is the corrected shelf life,where the wastage rate S≥ 0 B is the modified shelf life, where the loss rate S. The loss rate S can be obtained from annex IV, the loss rate of annex IV can be averaged to obtain the average loss rate of the sample, where B can be regarded as the corrected loss coefficient. The final results obtained see excel third ask process, after calculation, due to the small wrinkled skin (copies), green peduncle loose flowers, zhijiang green peduncle loose flowers of three types of single product profit is low, so it will be rounded off, rounded off after the remaining 33 single product of the total profit of 812.436.

Of these, cauliflower, eggplant, edible mushrooms, chili peppers, and foliage replenishments have not yet reached the total replenishment, and by scaling up the number of this single category in equal parts, we get a total profit of 1,175.539.

By collecting and analyzing various factors that affect the replenishment and pricing decisions of vegetable commodities. For example, data-driven decision-making, i.e., based on data analysis of various time periods, enables superstores to better understand the sales of vegetable commodities, market demand, etc., and make more informed replenishment and pricing decisions. Seasonal demand, i.e. different changes in sales and demand for vegetable commodities in different seasons. For example, some vegetables may be in short supply in a particular season, while they may be stagnant in other seasons. Costs and profits. When formulating pricing strategies, supermarkets assess the degree of market competition and price elasticity of vegetable commodities by taking into account production costs, storage costs and transportation costs, etc., to ensure that pricing covers costs and maintains a reasonable level of profit. Focusing on customer feedback and market trends, i.e. the superstore collects consumer feedback on the quality, selling price and taste of vegetable commodities, as well as competitors' sales strategies, so as to adjust the replenishment plan and pricing strategy of vegetable commodities. Through these uncertainties, the superstore collects and analyzes relevant data, establishes relevant models, and gives the corresponding replenishment and pricing decisions for vegetable commodities according to the future development trend. This enables supermarkets to have a more comprehensive understanding of market demand, inventory, market competition, and supply chain operations, so that they can more accurately formulate replenishment and pricing strategies for vegetable commodities.

5. Conclusion

This paper takes the measured data of a superstore as an example to formulate a set of effective pricing and replenishment decisions for vegetable products, and the comprehensive consideration ensures the balance of supply and demand, and reduces the loss of superstores and the loss rate of commodities. First of all, the sales of each category and single product in different time periods were counted, and the Pearson correlation coefficient was calculated to get the correlation intensity level and interrelationship between the representative eggplant single product, and the distribution law of the sales of each category and single product of vegetables was obtained by organizing and analyzing the data, so as to know the sales of each category of vegetables in different time periods, and the sales of different single products in the same category in different time periods. Popularity. Then, analyze the relationship between the total sales volume of each vegetable category and the cost-plus pricing, establish a regression model based on the measured data, and the results show that there is a regression relationship between the total sales volume and the cost-plus pricing, based on which, the pricing strategy is determined by the positive correlation between the ratio of the wholesale unit price and the sales unit price and the establishment of the Random Forest model to predict the total replenishment volume of the next week, and arrive at the pricing strategy of each category, and again Random Forest model is used to derive the pricing strategy for the next seven days. Finally, obtaining the replenishment volume and pricing strategy of individual items to maximize the revenue of the superstore, these individual items are scaled up in equal proportions, and the total profit can be calculated to be $1,175.539. The model established in the paper basically solves the given problem, combining strong practical strength and high computational efficiency.

References

[1] Huang, Jianxing. Research on Vegetable Price Fluctuation and Forecasting. South China Agricultural University, 2019.

[2] PENG Hongxing, ZHENG Kaihang, HUANG Guobin et al. Vegetable price prediction based on BP, LSTM and ARIMA models. Chinese Journal of Agricultural Mechanical Chemistry, 2020, 41(4): 193-199.

[3] Yan Zhengxu, Qin Chao, Song Gang. Random forest model stock price prediction based on Pearson feature selection. Computer Engineering and Applications, 2021, 57(15): 286-296.

[4] Cui, Yun-ho. Research on Fresh Vegetable Sales Prediction Based on CNN-PSO-LSTM Combined Model. Anhui Agricultural University, 2022.

[5] Lv Bin. Research on Vegetable Supply Chain Integration. Fujian Agriculture and Forestry University, 2010.

[6] Han, W.-G. Research on inventory management program of frozen products category in community vegetable direct stores in Urumqi. Xinjiang Agricultural University, 2017.